DOCUMENT RESUME

ED 432 025 EA 029 243

AUTHOR	Flanders, Anne K.; Wick, John
TITLE	Peer Evaluation of the School's Potential To Improve Learning.
PUB DATE	1998-04-16
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Accreditation (Institutions); *Educational Assessment; Educational Improvement; Elementary Secondary Education; *Evaluation Methods; *Interrater Reliability; Longitudinal Studies; *Peer Evaluation; Program Evaluation; *Program Validation
IDENTIFIERS	North Central Association of Colleges and Schools; *Outcomes Accreditation

ABSTRACT

This paper examines whether the peer-review process of the North Central Association (NCA) is reliable and valid. Reliance on peer judgments has been a part of NCA accreditation, but confidence in the use of peer decisions to certify a school's readiness to implement the improvement plan--Outcomes Accreditation (OA)--was weak. The study focused on three questions: Did the peer reviews reflect criteria for OA school improvement or other factors? Did the peer reviewers make accurate judgments even though improvement plans and school characteristics differed? and Can reviewer accuracy be predicted? The study drew on ratings and diagnostic feedback from 245 reviewers involved in OA peer review from 1992 through 1994. Most of the reviewers were school principals. More than 1,500 independent reviews of school-improvement plans were studied. The results indicate that reviewers were most accurate when they acted on well-developed beliefs specifically related to OA activity and then transferred those beliefs to the evaluation of another school's set of improvement goals. A variety of factors influenced individual accuracy, including the individual's engagement, external environmental pressure for school improvement, and opportunities for collaborative professional exchanges. Reviewers applied OA criteria in accurate holistic decisions even when the schools and their improvement plans differed. Contains 14 references. (RJM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 432 025

Peer Evaluation Of The School's Potential To Improve Learning

by
Anne K. Flanders and John Wick
Northwestern University
Evanston, Illinois

The use of the holistic decision as the basis for granting schools candidacy for Outcomes Accreditation (OA) was initiated by the North Central Association (NCA) in 1992. Members of the voluntary accrediting organization felt the provision of feedback to schools before they entered into the work of improvement was crucial. After all, schools needed some assurance their school improvement was not guided by faulty premises before investing three to five years of effort in implementation. In addition, the membership of the organization needed a guarantee that schools were addressing the criteria they had agreed on for granting OA. This criteria established seven loose guidelines for school improvement that integrated research on cognition and learning, school effectiveness, and school improvement processes.

The reliability and validity of averaged peer ratings from the review of school improvement plans were not confirmed by NCA during the first years of their use. As a consequence, many assumed that peer reviewers were unreliable and felt the stakes of school improvement too great to leave in their hands. After all, falsely positive ratings would erode the rigor of OA criteria which stress improving the quality of learning for all students in the school. On the other hand, unjustly negative peer reviews punished schools unnecessarily. These schools already had devoted a year to assessing student learning and planning for improvements.

EA 029 243

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
A. Flanders
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

In fact, this study found that the holistic decisions of peer reviewers were not reliable critiques of the presence of OA criteria in the goals of school improvement plans at review initiation. Likewise, schools' sets of goals also barely reflected criteria for improving the quality of student learning equitably at first. However, as understanding about using the criteria in goal statements was built into improvement plans, they were followed by peer judgments that also reflected NCA OA guidelines for school improvement. After three years, the criteria was used fully by schools planning for improvement and also applied accurately in peer reviews. Moreover, differences between schools and in the materials they submitted for review did not deter reviewers from generalizing OA criteria accurately in their decision making.

The Problem

Reliance on the judgments of peers has been a part of NCA accreditation processes for years. However, confidence in the use of the holistic decision to certify a school's readiness to implement a plan that would cause improved learning outcomes was weak. The perceived risk of relying on the judgments of peers made it necessary to determine if these raters' evaluations reflected the criteria accepted by the membership of the accrediting agency. This study was undertaken to determine if the judgments made by peer reviewers were reliable and valid. Three questions were developed specifically to determine the reliability and validity of these reviews:

- Did the peer reviews reflect criteria for OA school improvement or other factors?
- Did the peer reviewers make accurate judgments even though improvement plans and school characteristics differed?
- Can reviewer accuracy be predicted?

Perspectives

Outcomes Accreditation is based on the belief that schools improve individually, so schools' improvement plans were expected to vary (NCA 1992, Wick and Gose 1994). Because schools weren't expected to have the same plans, evaluations of readiness to implement school improvement had to adjust to accommodate these differences. The use of a holistic judgment made during a review of the school's improvement plan was chosen because it could provide flexibility and individual feedback to schools, in addition to setting a focus for schools and peer evaluators on important guidelines for school improvement (Wick and Gose 1994).

The holistic decision rating that resulted from a review of a school's improvement goals was patterned after alternative evaluations being used to assess students' writing and portfolios (Wick and Gose 1994). Many have recognized the advantage of this type of alternative assessment for lending itself to nonstandardized products (Banta et al 1996, Stiggins 1995). Most agree these assessments are more comprehensive, provide better feedback on individual performance, and focus individuals and organizations on important criteria than more conventional means of evaluating student achievement. In Kentucky, principals and teachers found alternative assessment was an effective inducement to change practice (Koretz 1996).

Alternative assessments also have drawbacks. Many feel their reliability is weak (Gaddy, Hall and Marzano 1996, Koretz et al 1995). Therefore, Koretz and colleagues (1995), and Banta and colleagues (1996) have recommended strongly that the judgments of evaluators should be standardized on commonly recognized criteria. Moreover, in use they have proven to be difficult and resource draining (Banta et al 1996, Kane 1995). So, Stecher and colleagues(1997) suggest careful

examination of quality, feasibility, and potential usefulness of alternative assessments before human and material resources are committed to their implementation.

Some of these issues can be reduced. Yates (1990) suggests assessor calibration to improve reliability. In support, Chang and colleagues (1994) have found evidence that reviewers set higher standards for teacher certification criteria when they fully understood criteria and lower standards when the criteria was unfamiliar to them. However, Koretz and colleagues (1995) point out that the reasons for low reliability do not lie with the rater alone. In their study of Vermont Portfolio Assessments they found that problems interpreting and applying unconventional or confusing criteria also contribute to problems with reliability.

Investigations into the reliability of peer review for admission to OA candidacy have begun. Wick and Gose (1994) found the agreement between the reviewer's paired holistic decisions of a school's set of goals high. According to their research, only twice in two years did reviewer pairs vary more than one point on their holistic decisions. Using 1992 and 1993 reviewer surveys, Flanders (1993) also discovered that almost all reviewers agreed strongly with criteria that school improvement goals focus on student learning, and measurable and equitable improvement. However, questions about the reliability of reviewers' judgments when compared to criteria for OA and the validity of their assessments of non-standardized plans remain unanswered.

Methods

Sample

The participants in OA peer review for admission to candidacy from 1992 through 1994 provided ratings and diagnostic feed back to the schools that were

used in this study. Two hundred and forty-five reviewers took part in the first three years of OA candidacy for peer review: 45 in 1992, 92 in 1993, and 103 in 1994. Random sampling was not used to select these reviewers, though school improvement plans were assigned to them randomly for rating. Almost three-quarters of the reviewers (73%) were chosen by NCA state directors. The remainder of the reviewers (27%) were walk-in volunteers from NCA member organizations attending the annual conference in Chicago.

Reviewers were usually school level employees. The majority worked as school principals (63%). The remainder came from district offices (24%), school boards, state departments of education (2%), or institutions of higher learning (8%). Forty-eight percent of all reviewers were employed in high schools or by high school districts. The rest worked in K-12 (kindergarten through high school) districts (33%), elementary schools (13%), or junior high or middle schools (6%). In 1992 60% of the reviewers were male, though half of all reviewers were female in both 1993 and 1994.

One hundred ninety out of the 245 reviewers reported that they had no experience making the holistic decision before participating in the Chicago review. The 1992 and 1993 reviewer cohorts commonly noted they no practice making the OA holistic decision even though training and materials were made available to them. By the annual review 1994, 33% of all reviewers surveyed said they had used the holistic decision for peer review with school improvement plans either in Chicago or in their home state.

Data Collection

In all, four sources were used to collect data for the study of holistic decisions and their agreement with criteria for OA candidacy, including:

- "The Report for Outcomes Accreditation Schools," submitted by all OA candidacy applicants

- reviewer registration sheets for the OA candidacy review session in Chicago
- the "Review Process for Schools Seeking OA Candidacy" sheet, which peer reviewers filled out for every school they evaluated.
- The "NCA Reviewer Surveys."

The materials used in this study for candidacy application, peer review, and reviewer surveys are linked closely in format and content to provide consistency. The review document, "Review Process for Schools Seeking OA Candidacy," was developed from criteria for compliance with OA found on the inside cover of the "Review Process for Schools Seeking OA Candidacy." To maintain consistency, reviewer surveys were constructed to reflect the holistic decision and diagnostics in the "Review Process for Schools Seeking OA Candidacy."

School Data

The school data for this study was taken from information supplied by the OA candidates on the "Report for Outcomes Accreditation Schools." Schools report their city, state, school enrollment, number of professionals employed (FTE), school level (grades included), and school governance (public or non-public) on this form. The school's improvement plan and its description of the steps leading up to the development of the improvement plan (commitment, resources assistance, school profile, selection of target areas based on data about student performance, and the school's set of improvement goals) also are included in the report.

"The Report for Outcomes Accreditation Schools" files were collected on site during the 1992 and 1993 reviews for OA candidacy in Chicago. The data obtained from them is fairly complete. The same is true of the information taken from the final reports submitted by the 1993 and 1994 OA schools receiving full accreditation. In 1994, a copy of all of the candidate goals for one state were

obtained from that state's NCA office. This state had approximately one-third of the total OA candidates in 1994.

The sections of the report that are used during the annual peer review for advancement to candidacy are: Phase I-school commitment to OA (provides discussion of site based decision making and district level support); Phase II, number 3-notation of the identity of the resource specialist and visiting team chair, including professional affiliation, address and a description of the assistance given the school; Phase III-school profile summary; and Phase IV-selection of appropriate target areas and target area goals.

Census data was used to generalize 1993 school findings to broader populations. This data was obtained from *The County and City Data Book 1994: A Statistical Abstract* and *The 1990 Census of Population and Housing Supplemental Reports, Metropolitan Areas as Defined by the Office of Management and Budget* published by the Bureau of the Census. Slater and Hall's *Places, Towns, and Townships: First Edition 1993* also served as a resource for population data. It includes city or town classification by population density (metropolitan, urban, suburban, rural), population, total school enrollment, percentage of students enrolled in public schools, and families living below the poverty level.

Peer Reviewer Information

Background information on the peer reviewers was obtained from the registration records for the annual regional OA candidacy review and from surveys returned by reviewers. Reviewers were assigned numbers at the time of registration that they recorded on each review form. These numbers were used to link individual reviews to a specific reviewer. The school OA candidacy reviews and reviewer information for 1992, 1993, and 1994 also were collected at the annual regional review in Chicago and are complete.

The peer reviewers were surveyed for general perceptions of OA processes and OA precandidacy school efforts. All reviewers received "NCA Reviewer Surveys" from 1992 through 1994. In 1992 these surveys were mailed to each participating reviewer; 61% were returned. In 1993 and 1994 reviewers were given surveys which were collected from them before they left the candidacy review. The 94% of the surveys were returned in 1993 and 86% in 1994.

Measures

Peer Review Based On A Fourpoint Forced Choice Scale

Admission to candidacy from 1992 through 1994 was granted when two peer reviews of the school's set of improvement goals had average ratings of 2.5 or more. Ratings are assigned by individual reviewers in response to the holistic decision question, "If a school improvement plan having this set of target area goals were faithfully implemented and learner outcomes were noticeably enhanced in these areas, to what extent would overall learner outcomes levels in the school have improved" (NCA peer review form). In addition, reviewers rated the school's plan on the seven diagnostics that are criteria for OA candidacy using the four-point scale. Some also added written comments to the school about their plan.

During a review of school materials the reviewer had to choose a rating level that implied the set of goals were either clearly acceptable (a rating of three or four) or unacceptable (a rating of one or two). In addition, because four rating levels were used, the reviewer indicated the degree of acceptability. In the peer review a rating of 4 signaled the goals were "Exemplary," 3 "Acceptable," 2 "Not Quite Acceptable," and 1 "Unacceptable." A rating of 4 indicates the reviewer believes the set of goals will have a significant impact in improving learner outcomes. On the other hand, a rating of 1 means the reviewer thinks the school's

goals will have no significant impact on improving learning in the school. A 3 denotes the reviewer's feeling that improvement will be fully acceptable, and a 2 that the school's plan will not make enough impact on student learning.

Each review session in Chicago began with a short lecture on OA criteria, models of goals and ratings, and a practice review. This was done to calibrate reviewers' holistic decisions before they began the actual reviews of school materials. Almost all reviewers reported that they found this session helpful. In 1994 reviewers worked cooperatively on practice reviews. In addition, they were allowed to cooperate as a team on one of the two independent reviews of the school's plan.

Analysis And Comparison Of The Contents Of OA Candidate Applicants' Sets Of Improvement Goals With Peer Reviews

In order to compare the schools' sets of goals to the holistic decisions they received from peer reviewers, it was necessary to build a means of measuring the degree to which goals met criteria for OA candidacy. To do this, contents of OA candidate sets of improvement goals were analyzed using a coding system developed from the diagnostic feedback found in the peer review form. The coding for each goal included: (1) the type of outcome expected (complex behavior, indicator, implementation of a process, or organizational), (2) the curricular or extracurricular target selected, (3) notations of evidence of decision made on student data and equal expectation for improvement by all students', and (4) the level of learning being addressed (skill and knowledge acquisition, or the integration and use of knowledge and skills in complex activities). The coding process was developed, tested, and refined over a two year period in a variety of settings with educators from the NCA region (Wick and Gose 1994).

The rating used here for the school's set of goals in its match to OA criteria for candidacy is the Match To Tempiate (MTT). The MTT was based on the assumption that each of the OA criteria could be indicated by constructs. Therefore, the percent of all goals in a school's plan that were a construct for each diagnostic category from the review form could, when averaged together, give a rough approximation of the extent to which the school's goals meet OA criteria. The constructs indicated were: (1) the goals are stated as student learning outcomes, (2) they are equitable, (3) their selection is based on student data, (4) they challenge the school's students, (5) they are focused on higher level skills, (6) they involve the school's staff, and (7) they are coordinated and integrated as a set.

One further assumption was made. The MTT of a school's plan converted to a fourpoint scale was assumed to provide a comparison between the school's compliance with OA and the reviewer's holistic decision. A sample determination of a school's MTT can be found in Appendix I. Four steps were required to determine the extent to which a set of school goals meets OA criteria. These were: (1) a content analysis of the goals in a school's OA plan to identify OA criteria in each goal statement; (2) a determination of the percent of goals in the plan that have each criteria, which are student learning outcomes, different complex behaviors, different curricular foci, addressing integrated learning, equitable, and databased; (3) an overall average for compliance determined by averaging the percentage of each criteria for the school's set of goals together; (4) conversion of the overall average from a hundredpoint scale to a fourpoint scale as used in the rating for the holistic decision

Procedures

More than 1500 independent reviews of school improvement plans were studied. Three separate data bases were developed to answer the three research

questions that are the focus of this study. One linked the contents of school improvement plans and school characteristics to the OA template using the MTT. Another linked the MTT of the schools to peer review ratings and reviewer characteristics. The final database provided more in-depth information about the schools and improvement plans of the top and bottom quartiles based on reviewer ratings in 1993.

Correlations were calculated the peer reviewers' individual and averaged holistic decisions and the MTT of the 1992, 1993, and 1994 holistic decisions they reviewed. Reviewer accuracy was identified using the absolute difference between the reviewer's averaged peer reviews and the average MTT of school plans he or she reviewed. Changes in the difference between MTT and holistic decisions were investigated through year-to-year comparisons. Controls were in place to assure random assignment to reviewers and to prevent any violation of the reviews' independence in making the holistic decision in 1993. Therefore, descriptions, comparisons, and inferences are made from data collected from the 1993 OA candidate upper and lower quartiles by averaged peer review.

Results

Did The Peer Reviews Reflect Criteria For OA School Improvement Or Other Factors?

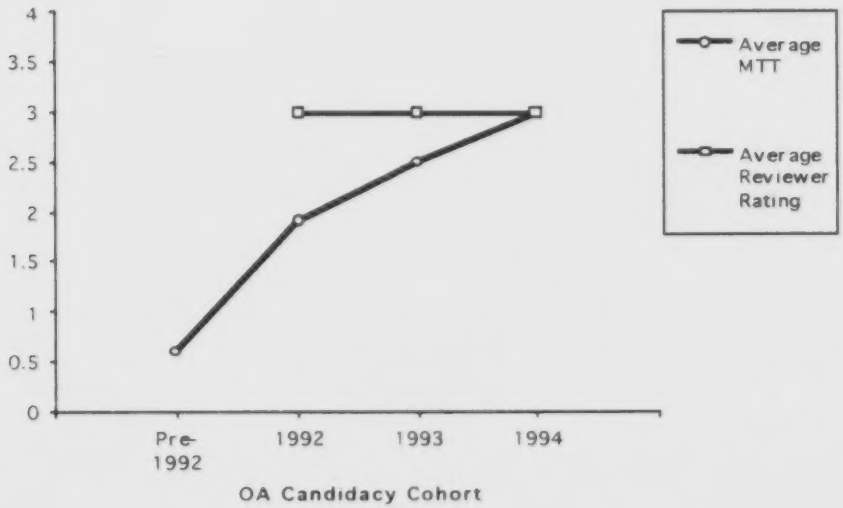
The averaged holistic decision rating remained the same during the three years of this investigation. In fact, averaged peer review ratings on the holistic decision varied insignificantly from 1992 through 1994 [F (2, 565) =.454, p > .05]. However stable the ratings, the gap between holistic decisions and the degree to which sets of school goals addressed OA criteria closed (see Figure 1).

As can be seen in Figure 1, on the average 1992 reviewers were more than one point apart from the degree of OA criteria evident in the candidate's set of goals

when assigning ratings. By 1993 the gap between the two closed to less than one-half point. In 1994, almost 75% of all reviewers differed less than one half a point on the average in their holistic decision ratings from the MTT of the schools they reviewed. During the first two years of the peer review, false positive decisions were more predominant than negative ones because peer ratings consistently exceeded the MTT.

Correlations calculated for average reviewer ratings (holistic decisions) and the MTT of the OA candidate plan the two peer reviewers evaluated were significant but weak (see Table 1). In general, higher ratings usually were given to candidates with higher matches to the OA template. Lower holistic decisions usually were given to sets of goals with a lower match to the OA template during all three years of this study.

Figure 1 Comparisons Between Averaged Holistic Decision Ratings And The Degree Of Match To Template (MTT) Of Sets Of School Improvement Goals



Though reviewers' holistic decisions were stable, the difference between them and the MTTs did decrease between 1992 and 1994.

Table 1 Correlations Between OA Candidate's MTT And The Averaged Holistic Decisions Given Schools' Sets Of Goals

	Degrees of freedom	*r*
1992	153	.25**
1993	315	.194**
1994	82	.394***

** *p* significant at .01, for a two-tailed test, *** *p* significant at .001, for a two-tailed test

It also was discovered that different characteristics of sets of school goals influenced holistic decisions from year to year. When regression was used to identify important criteria for holistic decision making (see Table 2), it can be see that reviewers focused on some aspects of the OA template and shifted their interest away from or ignored others. Only the number of goals (approximately four per plan) remained significant in reviewer's decision making processes during this investigation. The adjusted R^2 suggests that few of the characteristics listed here have any significance in reviewers' decisions in 1992 and 1993. From the regression equations, OA criteria appear to have little effect on holistic decision making. Even in 1994 it explained little more than 30% of the decisions reviewers made that year. Since reviewer accuracy is the basis for reliability in this type of assessment, it was examined to see if accurate and inaccurate reviewers made decisions differently.

Table 2 Impact Of Schools' Sets Of Improvement Goals On Averaged Decisions

Variable [a]	REGRESSION COEFFICIENTS (t ratio in parenthesis) 1992	1993	1994
Total number of goals	-0.107 (-2.227*)	-0.079 (-1.940*)	-0.397 (-3.961**)
Goals for different curricular areas	-0.018 (-0.432)	0.093 (2.957**)	0.308 (2.825**)
Goals that indicate learning	0.096 (2.250*)	-0.023 (-0.648)	0.036 (0.451)
Goals for learning processes	0.002 (-0.041)	0.047 (1.127)	-0.069 (-0.677)
Goals for complex behaviors	0.017 (0.276)	0.005 (0.207)	0.058 (0.641)
Goals for school outcomes	-0.021 (-0.027)	-0.263 (-3.123**)	-0.036 (-0.211)
Goals integrating learning	0.026 (0.543)	0.034 (1.205)	0.016 (0.212)
Measurable goals	0.011 (0.246)	-0.026 (-0.916)	0.081 (0.866)
Goals including all students	0.103 (2.367*)	0.035 (1.647)	0.302 (3.404**)
Intercept	3.160 (24.522)	2.821 (24.982)	2.254 (6.055)
Adjusted R^2	0.07	0.08	0.32
Standard Error	0.56	0.52	0.56
Sample size	164	310	84

* $p < .05$. for two-tailed test , ** $p < .01$ for a two-tailed test

a. All independent variables are counts of the number of goals in a school plan that fall in that category

Regression equations were employed to determine which OA criteria in sets of goals influenced the inaccurate and accurate decisions given schools in the first and fourth quartiles by averaged rating of the 1993 cohort (see Table 3). One hundred sixteen of the peer reviews were accurate; they had the same MTT and holistic decision rating. Fifty-one were considered inaccurate because their MTT and holistic decision ratings were two to three points apart on the fourpoint scale.

BEST COPY AVAILABLE

Use of OA criteria in sets of goals explained about 80% of accurate reviews. Only 42% of the decisions made by inaccurate reviewers could be accounted for by OA criteria, and the standard of error was high. As can be seen, equity in goal statements significantly lowered inaccurate decisions, and challenge significantly raised accurate decisions. However, higher level skill use was a significant factor in higher holistic decisions for both types of reviewers.

Table 3 Decisions From 1993 Fourth And First Quartiles By Averaged Holistic Decision

	Regression Coefficients (t ratios in parenthesis)	
Variables	Accurate	Inaccurate
Percentage of goals that are student learning outcomes	0.074 (1.55)	0.511 (1.61)
Percentage of goals that are equitable	0.082 (0.805)	-0.594 (-1.828*)
Percentage of goals that are selected from student data	0.028 (0.313)	-0.275 (-0.957)
Percentage of goals that are challenging to the school's students	0.354 (1.704*)	-0.234 (-0.312)
Percentage of goals that are outcomes for higher level skill use	0.147 (2.15**)	1.126 (2.02**)
Percentage of goals that involve school staff	0.153 (0.906)	0.291 (0.476)
Percentage of goals that are coordinated and integrated	0.068 (0.331)	-1.379 (-1.526)
Adjusted R^2	0.80	.42
Standard Error	0.33	.89
Sample size *	115	51

* $p < .10$, for two-tailed test ** $p < .05$, for two-tailed test *** $p < .01$ for a two-tailed test

Do Peer Reviewers Make Accurate Judgments Even Though Improvement Plans And School Characteristics Differ?

An attempt was made to determine if reviewers' holistic decisions were valid. In other words, did they truly reflect the contents of goal sets when measured against OA criteria? If not, did reviewers rate schools differently

BEST COPY AVAILABLE

depending on characteristics, environmental factors, or school's organization and implementation of OA pre-candidacy activities?

The first quartile (1.5 to 2.5 averaged rating) and fourth quartile (3.5 to 4 averaged rating) of 1993 OA candidates were compared based on school characteristics, environmental factors, and OA processes. When the holistic decisions from 1992 through 1994 were examined in conjunction with characteristics, no variables were found indicating that peer reviewers' holistic decision ratings correlated with differences in school size, staffing ratios, grade level, or governance (see Table 4).

Table 4 Individual Reviewers' Holistic Decisions And Characteristics Of Schools From The First And Fourth Quartiles By Averaged Holistic Decisions, 1993

School Variables	Correlation
State in NCA Region	-.1998* to .1586
School Governance	.0067
High School	-.1074
Middle School/Junior High	-.0169
Elementary School	-.1649
Enrollment	.1823
FTE	.1656
Pupils-per-Teacher	-.1199
degrees of freedom (*df*)	158

* $p \leq .05$ for a non-directional test; ** $p \leq .01$ for a nondirectional test

One variable for state correlated significantly with reviewers' lower holistic decisions

Comparisons of the environments (poverty rate, median income, population density, school age population) of these schools with holistic decisions indicated

these factors had no significance in reviewer ratings in 1993 (see Table 5). However, school level participation at the initiation of OA processes correlated significantly with higher peer ratings. In addition, higher holistic decisions also were found in conjunction with the involvement of state departments of education or private consultants in goal setting. Lower holistic decisions were associated with studies conducted by the school alone.

Table 5 Averaged Reviewers' Individual Holistic Decisions From In The First And Fourth Quartiles, 1993

SCHOOL ENVIRONMENT VARIABLES	*df*	Correlation
Local population	116	-.0536
Percent of local population that is minority	111	.0399
School aged population	80	-.0342
School located in rural area or small town	116	-.0534
School located in suburb	116	.0439
School located in city	116	.0995
School located in large metropolitan area	116	.1636
Percentage of schools in the area that are public	88	-.0488
Median family income	102	.0777
Percentage of families below the poverty level	113	.0443

* $p \leq .05$ for a non-directional test ** $p \leq .01$ for a nondirectional test

There are no significant correlations between school environment variables and reviewers' averaged holistic decisions. Statistics are based on 1990 Census data and the 1993 Census yearbook.

Table 6 Holistic Decisions With Organizational Participation From First And Fourth Quartiles, 1993

OA Process Step	Organizational Participants	Correlation
I. Initiation:	School alone	.2477*
	School and school district office	.1727
	School and Consultant	-.0269
II. Commitment:	School alone	.0597
	School district office alone	-.1679
	Consultant alone	-.0630
III. Self-Study:	School alone	-.1948*
	School and school district office	.0412
IV. Goal Setting:	School and State/Private Consultant	.3819**
	School Alone	.0963
	School and District Office	.0973
	State or Private Consultant	.1026
	degrees of freedom (df)	158

* $p \leq .05$ for a nondirectional test ** $p \leq .01$ for a nondirectional test

Schools' independent initiation of OA processes, and the participation of the state in school goal setting correlated significantly with higher holistic decisions. Lower holistic decisions correlated significantly with the school conducting its self-study without outside assistance.

Can Reviewer Accuracy Be Predicted?

Differences in the compliance of schools' sets of goals with OA criteria (MTT) emerged when comparisons of accurately and inaccurately rated sets of goals were made (see Table 7). Sets of goals receiving inaccurate first quartile averaged reviews (low ratings) were higher overall in their MTT than inaccurate fourth quartile reviews (the highest ratings). Inaccurately high peer reviews were lower on the percentage of goals that were equitable than the average 1993 OA candidate applicant. In fact, it appeared that inaccurate and accurate reviewers had opposite definitions about some aspects of OA criteria, especially equity, challenge, and staff involvement.

Table 7 Comparisons Between Accuracy In Holistic Decision Making And OA Criteria

	1993 OA CANDIDATES			
	Fourth Quartile		First Quartile	
OA Criteria	Accurate	Inaccurate	Accurate	Inaccurate
Learning focused	91%	61%	63%	59%
Equitable	93%	48%	32%	56%
Measurable	97%	70%	76%	70%
Challenging	85%	49%	51%	58%
Higher skills	80%	64%	62%	65%
Staff Involved	89%	60%	62%	72%
Coordinated and integrated	81%	65%	64%	68%
MTT	88%	55%	60%	65%

The MTT of inaccurately rated OA candidates was least in the fourth (top) quartile and greatest in the first (bottom) quartile.

Multiple regression analysis was conducted on the holistic decisions of all 200 reviewers to determine which, if any, of the variables in reviewers' backgrounds contributed to their accuracy. Occupation, level of schooling, state, and sex made no difference in holistic decisions. However, certain factors were significant for reviewer accuracy during the first three years the peer review was used. These were: (1) the level of OA activity in the reviewer's state, (2) the reviewer's engagement in acquiring knowledge about OA as reported on the annual reviewer survey, and (3) the extent to which candidates' sets of goals addressed the template. The adjusted R^2 suggests that a combination of these factors accounted for 46% of the degree of agreement between reviewer's average holistic decisions and a school's set of goals as matched the criteria for OA (see Table 8).

Table 8 Reviewer Accuracy In Holistic Decision Making

VARIABLE	Coefficient	Standard Error	*t* ratio
State level activity			
Previous reviewers [a]	0.002	0.001	1.686
Previous candidates [b]	0.0004	0.0004	1.175
Reviewers present [c]	-0.004	0.002	-2.543**
Invited Reviewer [d]	-0.032	0.016	-1.973*
Average school MTT	-0.298	0.061	-4.865**
Reviewer engagement			
Seeking OA Candidacy [e]	0.004	0.018	0.197
OA Workshop [f]	-0.423	0.022	-1.981*
NCA/OA print/media use [g]	-0.071	0.026	-2.770**
OA Personal contacts [h]	-0.080	0.030	-2.686**
Years holistic decision in use	-0.021	0.017	-1.201
Adjusted R^2	.46		
Sample size	200		

* $p \leq .05$, $-1.960 \leq t \geq 1.960$, two-tailed test ** $p \leq .01$, $-2.576 \leq t \geq 2.576$, two-tailed test

Negative *t* ratios indicate a decrease in the difference between holistic decision and the MTT of a school's set of goals. Reviewer accuracy appears to increase when the state and the reviewer are both actively involved in OA , and when schools develop understanding of integrating OA criteria in goal statements.

a. Number of peer reviewers coming from the reviewer's state prior to the current year.
b. Number of OA candidates coming from the reviewer's state prior to the current year.
c. Number of peer reviewers coming from the reviewer's state this year.
d. A dummy variable signifying whether the reviewer was invited to participate based on NCA state office recommendation.
e. A dummy variable indicating whether the reviewer identifies him or herself as either seeking or preparing to seek OA candidacy.
f. A dummy variable indicating whether the reviewer indicated that he or she attended an OA workshop during the past year.
g. A count of the types of print or video information sources reviewer reports using to get information about OA during the past year.
h. The number of the different types of personal contacts (state office, school visits, informal peer discussions, and conversations with peers seeking specific information) the reviewer reports using to obtain information about OA during the past year.

Conclusions

We conclude that applying a correct holistic decision to an OA applicant's set of goals and other materials submitted by a school was not an impossible task for these reviewers. However, it was a task that took learning and time before good decisions became the rule rather than the exception. Expertise in making correct holistic decisions did develop in peer reviewers as understanding of the criterias' use in school improvement goal statements grew.

Overall, the reviewers who were subjects of this study were most accurate when they acted on well developed beliefs specifically related to OA activity and then transferred those beliefs to the evaluation of another school's set of improvement goals. A variety of factors influenced individual accuracy, including the individual's engagement, external environmental pressure for school improvement, and opportunities for collaborative professional exchanges. In addition to identification by the NCA state office, these factors have potential use for identifying accuracy in reviewers to ensure greater reliability.

Reviewers applied OA criteria in accurate holistic decisions even when the schools and their improvement plans differed. This makes it clear, that the most significant aspect of decision making was the criteria identified for school plans by the accrediting organization. Holistic decisions did not vary significantly with school characteristics or characteristics of the school's environment. However, higher holistic decisions were associated with OA processes that involved schools in initiation, where school districts or outside consultants helped with self-study and when states or outside consultants assisted schools in setting goals. These factors might suggest that school willingness to improve and collaboration with external agencies or individuals have positive effects - at least as far as goal statements. However, this will have to be investigated further to be verified.

Accuracy in holistic decision making appears to follow the use of criteria in goal statements, not precede it. During the transformation of theory (OA criteria), into use, expertise in evaluation developed. At least initially, accuracy in holistic decision making did not seem to improve use of the criteria in sets of goals, it followed use of the criteria in goal statements. The evidence that most clearly suggests that schools' compliance with OA guidelines had the most effect on accurate holistic decisions is the growth in MTT from year to year, even when peer review feedback was inaccurate. The stability of the average peer review ratings also indicates that the accuracy of the reviewers may have been keeping pace with the development of knowledge about the use of criteria in the goal statements. Therefore, these judgments were accurate reflections of the understanding of the use of template criteria as far as it had developed at that point.

Assimilation of OA criteria was crucial for reviewer accuracy. Evaluator reliability took two years to develop. However, identification of OA criteria began in the middle 1980s, and some of it was being implemented by schools before it was formalized and used in the peer review. This leads us to believe that despite the stability of the averaged reviews, peer reviewers gradually integrated and applied the template components of OA in holistic decisions. In other words, they wait until improvement plans reached their level of rating. This leads us to believe that the peer review based on the holistic decision was well adapted to the initial implementation of school improvement guided by loose criteria. As it was the collective judgment of peer reviewers reflected as much of the criteria as was usable at the time. So, schools were not punished for neglecting criteria still under development.

In the first two years, false positive holistic decisions far outweighed false negative ones. Irregularities between reviewer ratings, even if only by one point

caused problems. They appeared most often when schools in the same district or from the same state submitted identical goals and received different ratings. This resulted in complaints and a loss of confidence in the peer review. Submission of identical plans does not communicate the spirit of attempting school-based improvement, but reviewers were not finely tuned, either.

Fortunately, false positive peer reviews did not deter development and use of the OA criteria. We tend to believe that the use of peer reviews may have been fairer than measurement through indicators. The MTT is an example of a rigid evaluation system based on indicators that could have been used to determine systematically if schools were ready for improvement. However, if the MTT had been used to qualify candidates for school improvement, more than 50% of the schools would have been barred from candidacy in 1992.

The holistic decisions of accurate reviewers were valid. Accurate reviewers focused their decision making on the OA criteria contents of goal sets of rather than school characteristics or other factors. In addition, there is little indication that inaccurate reviewers largely considered school characteristics in their decision making processes. Statistics for the environmental characteristics used in this study were obtained from Census data. In many cases they describe much larger populations than would be included in the school's attendance area. An investigation of characteristics based on school level data (median family income, poverty and minority levels, and school aged population) will need to be conducted to provide more conclusive evidence these factors were reflected in decision making.

Summary

Evaluator reliability facilitated by accuracy development is possible if the evaluator has opportunity and stimulus to incorporate a new belief system.

Organizations interested in improving schools through peer review need to offer evaluators opportunities learn about proposed change. Opportunities should include peer interchange as well as instruction, modeling and application. Likewise, persons who are working on making changes need to take advantage of these opportunities. Interchanges with peers, experience with making changes, informational media, and workshops or seminars are valuable resources for learning during change.

The subjective judgments of peer evaluators can be valid. Reviewers can focus on the assessment of the evidence of criteria rather than respond to personal beliefs or school characteristics. However, it takes time before judgments can be expected to give evidence of reliability and validity. Therefore, expectations for evaluation processes to cause change as well as expectations for change to occur should be tempered by the reality of implementing change.

The holistic decision assessing potential for school improvement is far more forgiving and just to schools that are in the midst of constructing change than a more mechanical system was. It is possible that an evaluative judgment, such as the holistic decision, is most appropriate for continual improvement. Human judgment is an assessment system that is capable of learning and refinement. Professional judgments studied in this research were reliant on common knowledge about criteria at that point. As knowledge about use of criteria evolved, so did the rigor of the judgments made by reviewers.

Persons and organizations that are using or contemplating the use of peer review may consider the findings in this study. Three recommendations we would make to those who want to promote improvement at the school level under a loose set of guidelines and evaluative judgements are:

- Do not expect rigorous evaluations based on criteria at the beginning of fundamental change processes when alternative assessments are used
- Do expect peer reviews to reflect criteria as it is used in the product being evaluated: and
- When it is important to assess the degree reviewer judgments match criteria, devise a measuring stick to approximate the amount of criteria evident in the products that are being evaluated. Random sampling products and their ratings should give a good indication of progress.

REFERENCES

Banta, Trudy, Jon Lund, Karen Black and Frances Oblander, *Assessment in Practice: Putting Principles to Work on College Campuses*. San Francisco: Jossey-Bass, 1996.

Chang, Lei and others, Does a Standard Reflect Minimal Competency of Examinees or Judge Competency? Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, April 4-8, 1994.

Department of Educational Statistics (DES) *The Digest of Education Statistics*. Washington, D.C.: US Department of Education: Office of Educational Research and Improvement, 1993 and 1996.

Flanders, Anne, *When Peers Rate School Improvement Plans*. Tempe: North Central Association Commission on Schools, 1993.

Gaddy, Barbara, T. William Hall and Robert Marzano, *School Wars: Resolving Conflicts Over Religion and Values*. San Francisco: Jossey: Bass 1996.

Kane, Michael and Nidhi Khattri, Assessment Reform: A Work in Progress. *Phi Delta Kappan*. September 1995:30-32.

Koretz, Daniel, Sheila Barron, Karen Mitchell and Brian Stecher, Perceived Effects of the Kentucky Instructional Results Informational System (KIRIS). Santa Monica: RAND 1995.

Koretz, Daniel, et al. TheVermont Portfolio Assessment Program: Findings and Implications. Santa Monica: RAND 1996.

North Central Association, *A Focus on Student Success: A Handbook for Schools Seeking Outcomes Accreditation* . Tempe: NCA Commission on Schools 1994.

Rothman, Robert, *Measuring Up: Standards, Assessment, and School Reform*. San Francisco: Jossey-Bass 1995.

Stecher, Brian and others, Using Alternative Assessments in Vocational Education. Santa Monica: RAND 1997.

Stiggins, Richard, Assessment Literacy for the 21st Century. *Phi Delta Kappan*. November 1995: 238-245.

Wick, John and Ken Gose. *Improving Student Performance in Your School*. Dubuque, IA: Kendall/Hunt 1994.

Yates, J. Frank.*Judgment and Decision Making*. Engelwood Cliffs, NJ: Prentice Hall, 1990.

Appendix I

MATCH TO TEMPLATE (MTT) OF A SCHOOL'S SET OF GOALS

A "fit to template" is used to assess all OA candidate sets of goals for compliance with OA and to compare reviewers' independent judgments and the averaged peer reviews to the OA diagnostic template. An example of a school's OA improvement plan is used here. It would receive a rating of 70% for MTT, or 2.8 on conversion to a four-point scale. Table 1A provides an analysis of the plan by the goals included in it. It forms the basis of the score for the degree of MTT. Table 1B, which follows, provides a summary of the content analysis of the goals. This summary is applied to the diagnostics by assessing the extent to which the goals in the plan address the constructs for each diagnostic criteria.

SAMPLE SET OF OUTCOMES ACCREDITATION
SCHOOL IMPROVEMENT GOALS

I. All students will increase their respect for others and for property.

II. Students in speech class will demonstrate their ability to take a topic, choose and organize related ideas, and present their ideas clearly in standard English for the purpose of speaking to a group.

III. Students will demonstrate an improvement in test-taking skills.

IV. Students will use mathematical and scientific concepts in all curricular areas.

V. Students will develop the critical thinking skills necessary to develop solutions for problems in various mathematical settings.

Table 1A Content Analysis Of Individual Goals To Build A MTT

GOAL	Learning outcome	Complex behavior	Focus	Level of learning	Data-based	Equity
I.	yes	Caring for Self & Others	Citizenship	integrated use	yes	yes
II.	yes	Communicating	Speaking	acquisition	yes	no
III.	yes	Problem Solving	Study Skills	acquisition	yes	unsure
IV.	yes	Problem Solving	Problem Solving/ Thinking	integrated	yes	yes
V.	yes	Problem Solving	Problem Solving Skill	acquisition	yes	unsure

Table 1B Summary Analysis Of The School's Goals As An Entire Set

	Number	Percent
Total number of goals in plan	5	100%
Number of SLO goals	5	100%
Number of different complex behaviors	3	60%
Number of different curricular focuses	4	80%
Number of goals that integrate learning	2	40%
Number of goals that can be documented by data	5	100%
Number of goals that are clearly equitable	2	40%

From information in Table 1B, a MTT score for the entire set of goals is developed in the following manner:

1. SLOs: 100%
2. Equity: 40%

3. Data-based: 100% = (100% + 100%)/2 or (% SLOs + % measurable)/2

4. Challenge: 80% = (100% + 100% + 40%)/3 or (% SLOs + % measurable + % integrated learning)/3

5. Higher level skills: 73% = (100% + 80% + 40%)/3 or (% SLOs + % different complex behaviors + % integrated learning)/3

6. Staff involvement: 53% = (40% + 40% + 80%/3) or (% different curricular foci uses + % integrated learning + % equitable)/3

7. Coordination and integration: 47% = (60% + 40% + 40%)/3 or (% different complex behaviors + % integrated learning + % equitable)/3

Finally, the degree to which each construct is met is divided by the total number of constructs for the OA diagnostics (seven), resulting in a score for MTT; or the degree to which this school's plan matches the OA template. The resulting overall MTT of the sample set of school improvement goals used in the example is 0.70 or 70%. To convert this score to a four-point scale, like the holistic decision rating, it is multiplied by 4, equaling 2.8. The process for this particular set of goals can be condensed into:

$$MTT = (|100 + 40 + 100 + 80 + 73 + 53 + 47|/7)x4$$

In summary, the MTT is derived by going through:
a content analysis of the goals in a school's OA plan (Table 1A); determining the number and percentage of goals in the plan that are SLOs, complex behaviors, curricular foci uses, integrated learning factors, suggest equitable expectations for students, and include data-based element (Table 1B); and determining the percentage to which the set of goals meets the OA template diagnostics converted to a four-point scale.